Multilayer Scattering Image Analysis Fits fMRI Activity in Visual Areas

Michael Eickenberg, Alexandre Gramfort, Bertrand Thirion

Scattering Transform Layer One Linearizes functional MRI activation in Visual Areas
ICML Workshop on Statistics, Machine Learning and Neuroscience 2012

Michael Eickenberg, Alexandre Gramfort, Bertrand Thirion MICHAEL.EICKENBER@NSUP.ORG
Inria Parietal, Neurospin, CEA Saclay, Gif-sur-Yvette, France

Abstract

The scattering transform is a relatively new multi-layer and multi-scale signal transformation constructed to be invariant or tolerant to chosen transformations of the signal. Here it is used in its simplest form to fit activation in visual areas related to the presentation of natural images. As it performs a spatial pooling operation over high-frequency phase invariant edge detectors, it mimics the spatio-temporal low-pass filter properties of the haemodynamic response oberved in functional Magnetic Resonance Imaging (fMRI).

1. Introduction

Only recently have biologically plausible models of primate vision become successful enough to be competitive with state of the art methods in the domain of object recognition. In order to address tractability problems, they are mostly inspired by Hubel and Wiesel's classic result (Hubel & Wiesel, 1968) that the lowest level of processing is edge detection, the finding that visual cortex is hierarchically organized (Felleman & van Essen, 1991) and the fact that invariances and tolerances to certain transformations need to be built bit by bit along the hierarchy such that objects can be recognized despite differences in position, scale, orientation, illumination and pose. These ideas can be hard-coded into systems that implement the functionality of large groups of neurons and thus reduce the number of degrees of freedom left to be fitted to data if necessary. A common expression of this approach is the convolutional model, which alternatingly passes a linear filter (e.g. a linear edge detector, such as a Gabor filter) and a non-linearity over the input. In its 2007 version (Serre et al., 2007), the HMAX model achieves state of the art object recognition performance and human-like performance and errors in rapid object recognition tasks. It is implemented as an alternation of linear operations and pooling over space and scale using a maximum operation. The convolutional neural networks in (LeCun & Bengio, 1995) achieved state of the art performance on the recognition of hand-written digits.

Compared to electrophysiology, fMRI is a relatively new method to localize brain function. Albeit orders of magnitude coarser in spatio-temporal resolution and indirect in the nature of the signal, it permits the acquisition of the activity pattern of all the visual (and other) regions of a brain while a subject is viewing an image. Recently, fMRI brain activity of subjects viewing natural images has been shown to provide enough information to be able to identify the stimulus among a very large set of images (Kay et al., 2008) using an encoding model, and even perform a reconstruction of the stimulus using a "bag-of-images" prior and Bayes' theorem (Naselaris et al., 2009).

The advances in convolutional models for object recognition have led to an increased interest in their mathematical properties. Some degree of invariance is one of them and constitutes a key property of the signal processing in the brain. The *scattering transform* (Bruna & Mallat, 2011; Mallat, 2011) is a recent development and is built to create invariance to any chosen transformation group in a mathematically rigorous way. It consists of an alternation of an analytic directional wavelet transform and complex modulus followed by smoothing. The complex modulus is a non-linear component which adds stability by losing the phase of the transformation and prevents the smoothing from losing information (by averaging phase dependent edge responses). It turns out that the local averaging properties of the scattering transform may be well reflected

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

by fMRI voxels.

In this paper, the scattering transform will be briefly introduced and the analogy between its first layer and visual fMRI voxel properties will be established. The first scattering layer will be compared with the encoding model of (Kay et al., 2008), which consists of Gabor filter energies, and contrasted in predictive power by cross-validation.

Notation

For $u \in \mathbb{C}$ the complex modulus is noted $|u|$. Real signals are noted f and wavelets are ψ with scaling function ϕ.

2. The scattering transform

The scattering transform, introduced in (Mallat, 2011), is a signal transformation based on analytic wavelets which is constructed with the goal to establish a representation invariant or tolerant to any chosen transformation group on the signal while keeping its essential characteristics. By default, in images, the transformation group is translations and the essential characteristics are texture or object features. This property is enforced using a cascade of directional complex wavelet transforms followed by a complex modulus. At each stage, a smoothed (low-pass filtered) version of the intermediary result is sent to output, while the high frequencies are kept by the subsequent wavelet transform. The complex modulus acts as a convolution in Fourier space, which picks up correlations at a given frequency *difference*, thus bringing high frequency interactions within the Fourier support of the wavelet down to their frequency difference on each layer.

Here smoothing will be implemented as a local spatial average, building local translation tolerance.

Let Γ be a set of angles and $\psi_\gamma,\ \gamma \in \Gamma$ a directional analytic wavelet, e.g. a Gabor filter. With ϕ_J as the scaling function, we scale for $j = 0$ to $j = J$ as follows: $\psi_{\gamma,j}(x) = \frac{1}{2^j}\psi_\gamma(2^{-j}x)$. The first layer of the scattering transform of signal f is then

$$W_{\gamma,j}f \quad = \quad |\psi_{\gamma,j} * f|$$

The values of the first layer are smoothed

$$S_{\gamma,j}f \quad = \quad |\psi_{\gamma,j} * f| * \phi_J$$

and are output as such or globally averaged. The second layer and its output are as follows

$$\begin{aligned} W_{\gamma_1,j_1,\gamma_2,j_2}f &= ||\psi_{\gamma_1,j_1} * f| * \psi_{\gamma_2,j_2}| \\ S_{\gamma_1,j_1,\gamma_2,j_2}f &= ||\psi_{\gamma_1,j_1} * f| * \psi_{\gamma_2,j_2}| * \phi_J \end{aligned}$$

This can be continued and is further explained in (Mallat, 2011). The signal energy does not diverge and for appropriate wavelets the transformation is unitary. In analogy to the outputs of layers 1 and 2, the 0th layer output reads as follows:

$$S_0 \quad = \quad f * \phi_J$$

2.1. Scattering layer 1 and fMRI voxels

The early visual area V1 contains simple and complex cells that perform phase dependent and phase invariant edge detection. A contour can be defined by its location in the visual field, its orientation and its spatial frequency. A large number of these detectors is arranged in a retinotopic manner across V1. Retinotopy preserves local neighbourhoods, thus neurons that are close to each other will receive visual information from nearby positions in the visual field. This means that one fMRI voxel in V1 contains many neurons tuned to the same region in the visual field, while orientations and scales can be different. The hemodynamic response measured with fMRI can be seen as a spatiotemporal low pass filter. Since all the neurons inside the voxel contribute to its response, we can attempt to model it by locally averaging complex wavelet moduli of all scales. This method keeps high frequency energy, but blurs it out to the scale of the voxel. In fact, it can be viewed the first layer of the scattering transform. A similar principle is described in the DAISY transform (Tola et al., 2007) - a method for extracting scale invariant features that differs from the classic SIFT (Lowe, 2004) in that it takes local directional derivative averages instead of histograms of derivative orientation.

3. Methods

We use Morlet filters, i.e. Gabor filters with 0 DC offset in the *cos* part. This is achieved by choosing $C > 0$ such that the integral over

$$\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)(\cos(k^Tx) - C)$$

is zero. We use these filters throughout the experiment in order to permit a comparison between the two transforms.

We compared the performance of the first layer of the scattering transform on the (Kay et al., 2008) dataset

with the original filter pyramid that was employed. To this end, we implemented the pyramid according to the information provided in the supplementary material to the article: At 8 orientations and 5 scales, *cos* and *sin* type real, isotropic Gabor-like Morlet filters are placed at discrete locations on the stimulus image as follows: The middle of the image for the largest scale, at a spatial frequency of one cycle per image size. For the next scale the middles of the four sub-images of half the side length obtained by cutting through the middle of the image are occupied by filters one octave smaller, at two cycles per image size. The next scale is obtained by cutting each of the previous sub-images into 4 sub-sub-images and placing a filter in their center points. This process is continued twice more in order to end up with 256 filter locations on the smallest scale. The *cos* and *sin* parts of the filter response are combined by using the square root of the sum of squares, leading to a complex wavelet modulus as the transformation output. Adjustment parameters were the number of scales, J, chosen from $\{3, 4, 5, 6\}$, and the size of the first wavelet, determining the size of all the wavelets in the pyramid, chosen between a scale factor $\lambda_s \in \{\frac{3}{4}, 1, \frac{5}{4}\}$ times the original size.

Using the same type of wavelet, the scattering transform was performed using $J \in \{3, 4, 5, 6\}$ scales and 8 orientations. The output smoothing is defined by the largest scale through the convolution with ϕ_J. Hence, the size of this low-pass filter must be in accordance with the size of the receptive fields of the voxels, and especially not too large.

The dataset of Kay et al. 2008 consists of 1750 training stimulus images and 120 test stimulus images of size 128x128 pixels along with preprocessed fMRI voxel responses for two subjects of around 25000 voxels for each image. Minimally preprocessed fMRI data is also provided, but here we use only the preprocessed responses from both subjects.

In order to evaluate the linear predictive performance of the transformations, we perform a ridge regression on the transformation output as the data and the voxel response as the target value. We use 5-fold nested 5-fold cross-validation on the training set and predictive r^2 as the scoring function. The test set is not used for evaluation, as it is too small for our purposes and bears the risk of overfitting when used for parameter setting.

4. Results

For the Gabor pyramid the optimal settings measured by predictive r^2 are $J = 5$ scales and scale factor $\lambda_s =$ 1, as used in the original paper. A 6th scale adds 8192 coefficients to the 2728 of the setting $J = 5$, thus adding a lot of data which does not seem to be relevant to voxel prediction. On the other hand, $J = 4$ and less provides too few and too coarsely resolved coefficients. Hence we restrict ourselves to comparing the results obtained with these optimal settings to those of the scattering transform.

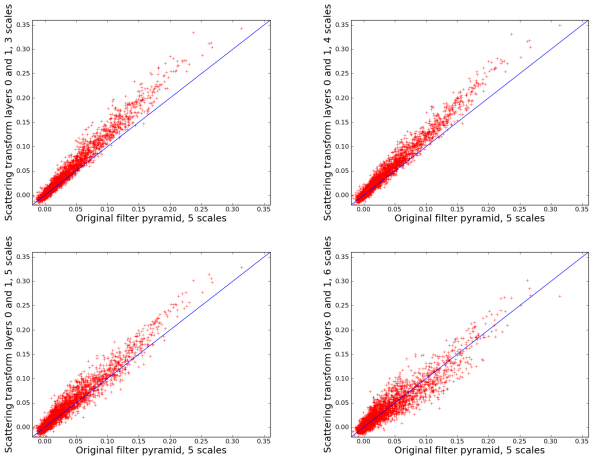

Figure 1. Scatterplots comparing predictive r^2 scores on subject 1. Each cross represents a voxel. The score using the original filter pyramid is on the x-axis. The y-axes are the scores using the scattering transform with different maximal scales J. Top left: $J = 3$, top right: $J = 4$, bottom left: $J = 5$, bottom right: $J = 6$. At higher scales the output is too coarse to fit voxel activity well. Scale $J = 4$ has the best results. Scale $J = 3$ has slightly worse results, possibly due to a higher number of coefficients in the regression.

All figures show data from subject 1, but the presented results equally hold true for subject 2. In Figure 1 we provide scatterplots comparing the linear predictive performance of the original model (x-axis) to the scattering transform first layer (y-axis) at different numbers of scales. For 3 and 4 scales the point cloud clearly lies above the diagonal. A Wilcoxon signed rank test confirms this with $p \approx 10^{-120}$ and $p \approx 10^{-200}$ respectively. The scattering transforms using 5 and 6 scales are nevertheless comparable to the original filter pyramid. In these cases, the output smoothing mechanism yields a response image size of 8x8, respectively 4x4 pixels per scale and orientation, which is too coarse to be able to resolve the receptive fields of typical voxels. As few as 3 scales in Fig. 1 top left are sufficient for the scattering transform to fit the voxel responses very well. In this case the output images are smoothed and downsampled to 16x16 pixels and are in accordance with the size of voxel receptive fields. This shows that

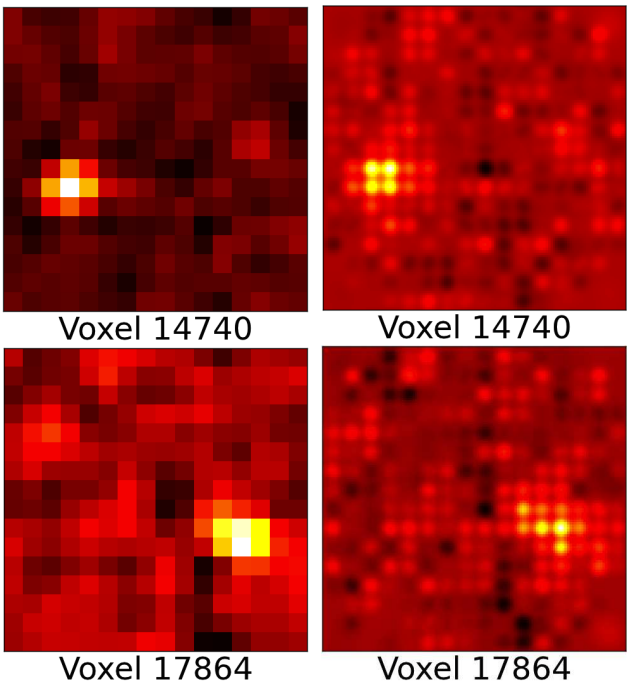

Figure 2. Receptive fields of two selected voxels using $J = 4$ scattering on the left and the original filter pyramid on the right. Receptive field size is on the order of $\frac{1}{8}$ of the side length. At least $\frac{1}{16}$ of the side length is needed to resolve it well.

the voxel activity can be fitted to edge filter responses of very fine scale, pooled over the region of its receptive field.

Figure 2 shows the receptive fields of two selected voxels for subject 1. They are visualized using different techniques: On the left there are the receptive fields due to the scattering transform, generated by taking the mean of the response over all scales and orientations. On the right, in order to visualize a receptive field using the original filter pyramid, we multiplied each resulting coefficient by the envelope of the corresponding filter. The discrete spatial nature of the pyramid becomes clearly visible. Both methods localize the receptive field in the same area and the signal energy is equally distributed over all orientations (not shown).

5. Discussion

In this paper we introduced the scattering transform as a candidate to model visual voxel activity. We compared it to the original Gabor pyramid and showed that fitting the size of a pooling region to the size of a voxel's receptive field yields higher predictive r^2 scores than fitting the size of the filter to the size of the voxel's receptive field. This gives rise to the hope that the subvoxel activity, which is composed of many visual neurons with receptive fields smaller than the one of the voxel, can be captured by attempting to model them and then pooling the result onto voxel size. Evidently, a rise in predictive score cannot be taken as evidence of having found a better model, but nevertheless, further inquiry into modeling of subvoxel activity followed by pooling as a means to model fMRI activity patterns seems called for.

Acknowledgments

References

Bruna, Joan and Mallat, Stéphane. Classification with scattering operators. In *CVPR*, pp. 1561–1566. IEEE, 2011.

Felleman, Daniel J. and van Essen, David C. Distributed hierarchical processing in the primate visual cortex. *Cerebral Cortex*, 1991.

Hubel and Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 1968.

Kay, Kendrick N, Naselaris, Thomas, Prenger, Ryan J, and Gallant, Jack L. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, Mar 2008. doi: 10.1038/nature06713. URL http://dx.doi.org/10.1038/nature06713.

LeCun, Y. and Bengio, Y. Convolutional networks for images, speech, and time-series. In Arbib, M. A. (ed.), *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.

Lowe, David G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pp. 91–110, 2004.

Mallat, Stéphane. Group invariant scattering. *to appear in Communications in Pure and Applied Mathematics*, 2011.

Naselaris, Thomas, Prenger, Ryan J., Kay, Kendrick N., Oliver, Michael, and Gallant, Jack L. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63:902–915, Sept. 2009.

Serre, Thomas, Wolf, Lior, Bileschi, Stanley, Riesenhuber, Maximilian, and Poggio, Tomaso. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:411–426, 2007.

Tola, Engin, Lepetit, Vincent, and Fua, Pascal. A fast local descriptor for dense matching. *Technical Report École Polytechnique Fédérale de Lausanne, CVLAB*, 2007.